Experimental design and statistical analysis

Michela Traglia and Reuben Thomas

Bioinformatics Core, GIDB
Gladstone Institutes

November 30, 2021

Introductions

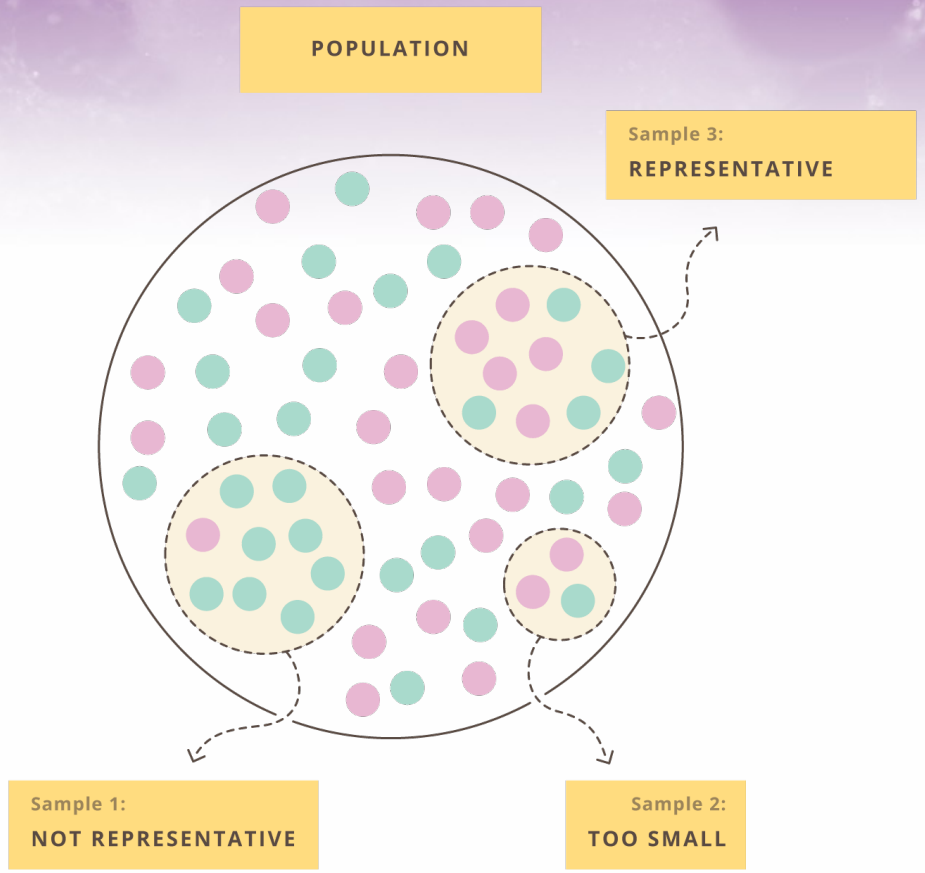
Michela Traglia

Statistician III

Reuben Thomas

Associate Core Director

Statistics is defined as
the science of collecting, analyzing, and drawing
conclusions from data



Generalization

- Empirical data are noisy
- Resources are limited

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research

Citation: Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2(8): e124.

Medical research has a credibility problem

- Estimated that ~75% published research findings cannot be reproduced
- ~$28 billion per year (nearly half of the annual non-clinical research budget in the US) is wasted on attempts to reproduce published studies
- Only a small percentage are due to overt fraud (intentional fabrication)
- Most are what are considered “detrimental research practices”
- Patient lives placed at risk

Thanks: Kevin Mullane
Director, Corporate Liaison & Ventures
Corporate Ventures and Translation
Gladstone Institutes

https://gladstone.org/events?series=responsible-conduct-of-research
https://rcr.ucsf.edu/

The purpose of the experimental design is to plan the experiment in a way that makes sure it can answer your biological question

Old journalists requirements of 5 Ws and 1 H
who, what, when, where, why and how

Motivation of the workshop

- Accessible statistical tools allow researcher to easily perform analyses
- How the program work and which setting to use?
- How to interpret the results?
- Hard to get defensible conclusions
- Before the statistical analysis, fundamental is to plan the experimental design
- Understand and critically evaluate scientific publications

Goals of this workshop

- Introduce basic concepts underpinning experimental design
- Learn how to think critically about the data you want to generate and use to make claims about
- Overview of statistical tests and concepts

Basic level course - No prerequisites

Please feel free to interrupt with questions, speak up or use the chat!

Real data is complex

- This workshop provides an introduction to typical experimental design and statistics.
- Experimental design can be very complicated
- Real data might need additional analyses choices.
- Some analysis choices need experience.

Consult with the Gladstone Bioinformatics core for such scenarios and data.

Outline

Experimental design principles

- Types of analytical studies
- Independent variable / response variable
- Target population and generalization

Statistical principles

Experiment: Assess differences between gene expression at two developmental stages

- Aggregation and inter-comparison
- P-value and multiple test correction
- Biological and technical replicates
- Power and sample size
- Overview statistical tests based on response and predictor

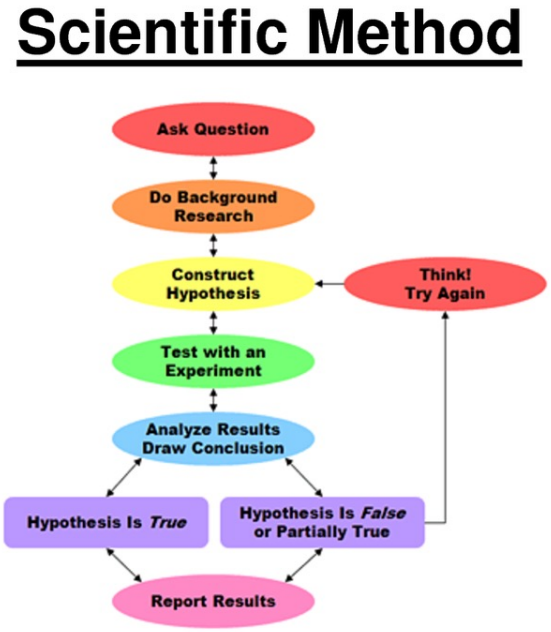
Confounding factors

Good vs bad experimental design

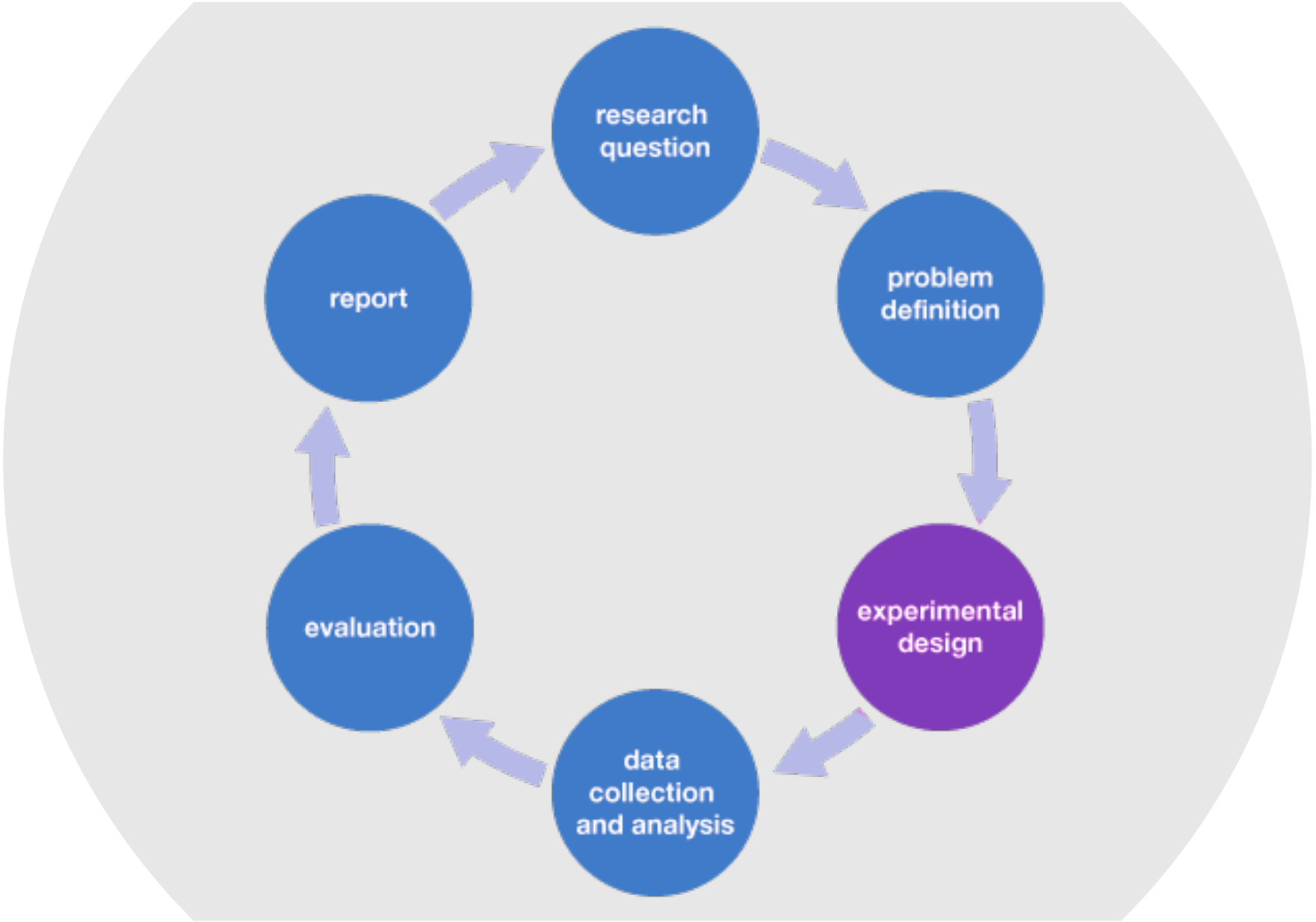
Statistical experimental designs

The scientific method consists of iterative application of the following steps:

(1) observing of the state of nature
(2) hypothesizing the mechanism for what has been observed
(3) collecting data
(4) analyzing the data to confirm or reject the hypothesis



Statistical experimental designs provide a plan for collecting data in a way that they can be analyzed statistically to corroborate the conjecture in question.


research question
problem definition
experimental design
data collection and analysis
evaluation
report

Ronald Fisher

1920 CE, Design of Experiments

Overcome the large amount of variation in agricultural and biological experiments that often confused the results

This motivated him to find experimental techniques that could

- eliminate as much of the natural variation as possible
- prevent unremoved variation from confusing or biasing the effects being tested
- detect cause and effect with the minimal amount of experimental effort necessary - time consuming and costly

Problem definition and research question

A specific issue, contradiction between two or more perspectives, or a gap in knowledge that you will aim to address in your research.

What is a (or the) scientific question that you are currently working on?

Analytical studies – define objectives

Observational studies	Experimental studies
Researchers don't assign choices	Researchers manipulate factors
Observing what is already happening	Create a treatment and compare the response
No establishing cause-effect	Changes cause an effect
Ex. Case-controls, cohort	Ex.Clinical trials

- Why is the experiment to be performed?
- Aim is to classify sources of variability? Or to study cause and effect relationship?

Problem definition and research question

Chemical pollutants during pregnancy contributing to autism in children

Lyall *et al,* Environ Health Perspect 2017

Poll

We want to study a potential association between pollutants measured in a cohort of pregnant women and autism diagnosed in their children (age 2-3).

Which kind of study are you dealing with?

1. Observational study
2. Experimental study

Independent variable and Response

Independent variable (IV)

Also called:

- Exposure variable
- Control variable
- Explanatory variable
- Manipulated variable

Dependent variable (DV)

Also called:

- Outcome variable
- Controlled variable
- Explained variable
- Response variable

Risk factors

Genotype

X

Disease

Gene expression

Y

In an experiment, you manipulate an independent variable to study its effects on a dependent variable (response)

Image source: by Hannah Bonville

Experimental units

Experimental materials to which we apply treatment and on which we make observations

- Animal or human subject
- Raw material for some processing operation
- Condition that exist at a point in time or trial?

Target population - generalization

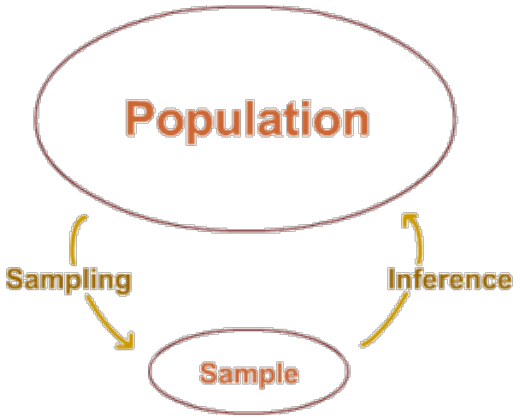
All subjects/units that we want base our claims/conclusions on

The cardiac tissue of all mice at embryonic stage E9.5

All children below 5 years old who are diagnosed with autism

All mother that were pregnant in a geographical area

Data is expensive
Studying of the sample ➔ Conclusion on the population



What are the conclusions you would like to draw from the results of your experiment?

Biological aspect to consider

- Does the samples reflect the hypothesis/question?
- What else is known beforehand on the topic? (e.g Expression levels on known interesting genes available?)
- What samples are available? More samples at a later time?
- Does the cell type express the genes of interest?
- All biopsies should be taken from the same part of the tissue!
- Do you have enough RNA from each sample or is pooling of samples required?

Outline

Experimental design principles

- Types of analytical studies
- Independent variable / response variable
- Target population and generalization

Statistical principles

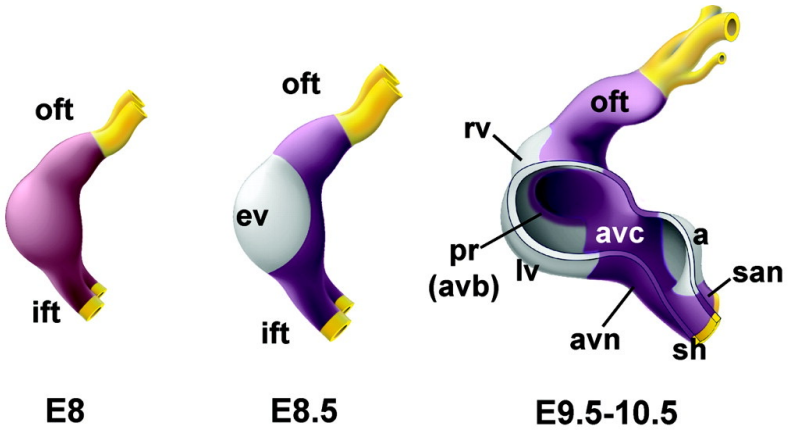
Experiment: Assess differences between gene expression at two developmental stages

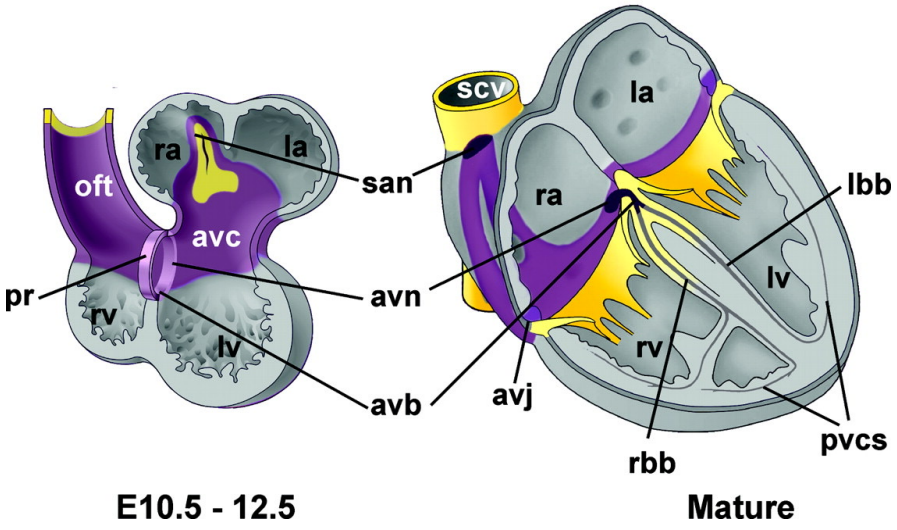
- Aggregation and inter-comparison
- P-value and multiple test correction
- Biological and technical replicates
- Power and sample size
- Overview statistical tests based on response and predictor

Confounding factors

Good vs bad experimental design

Problem definition and research question

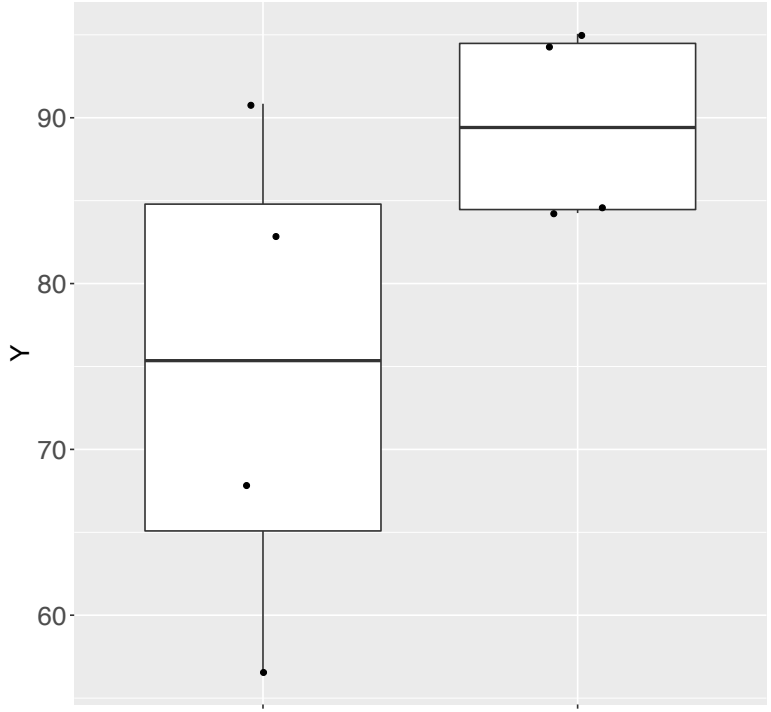


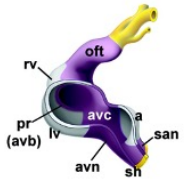


Gene controlling developing heart

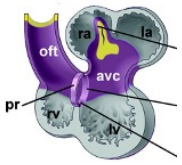
https://www.ahajournals.org/doi/epub/10.1161/CIRCEP.108.829341

Is gene differentially expressed between the two developmental time-points?



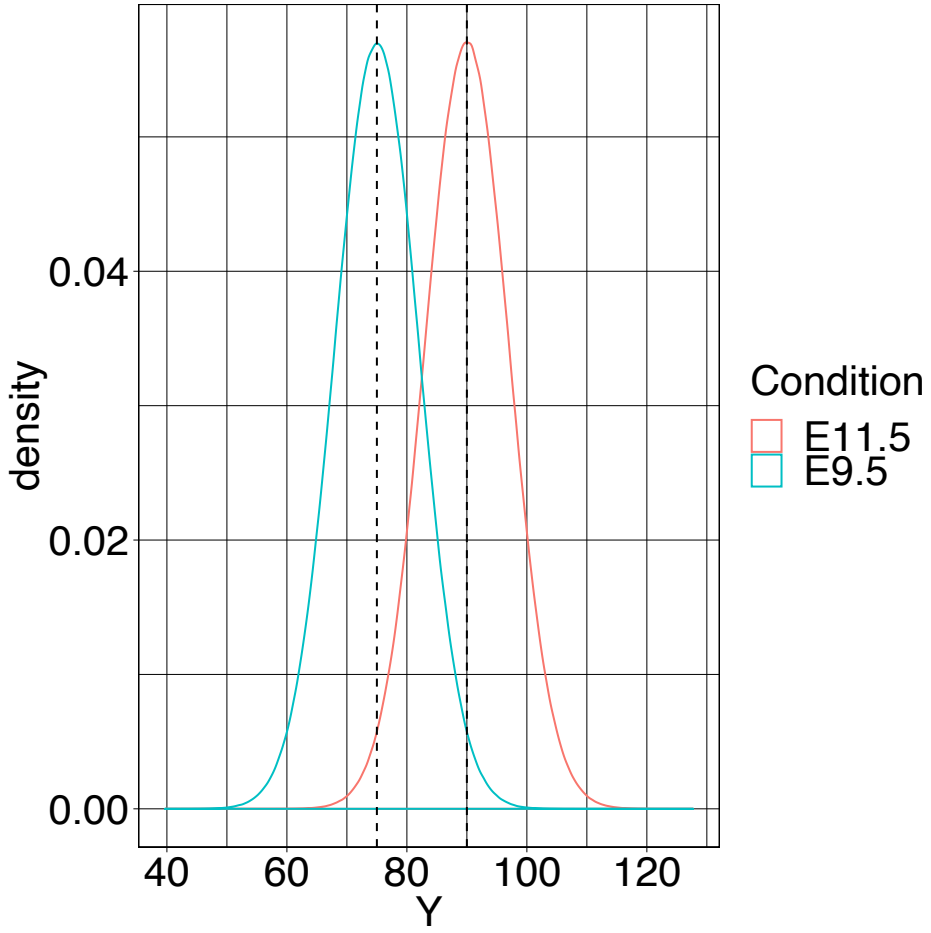


E9.5-10.5

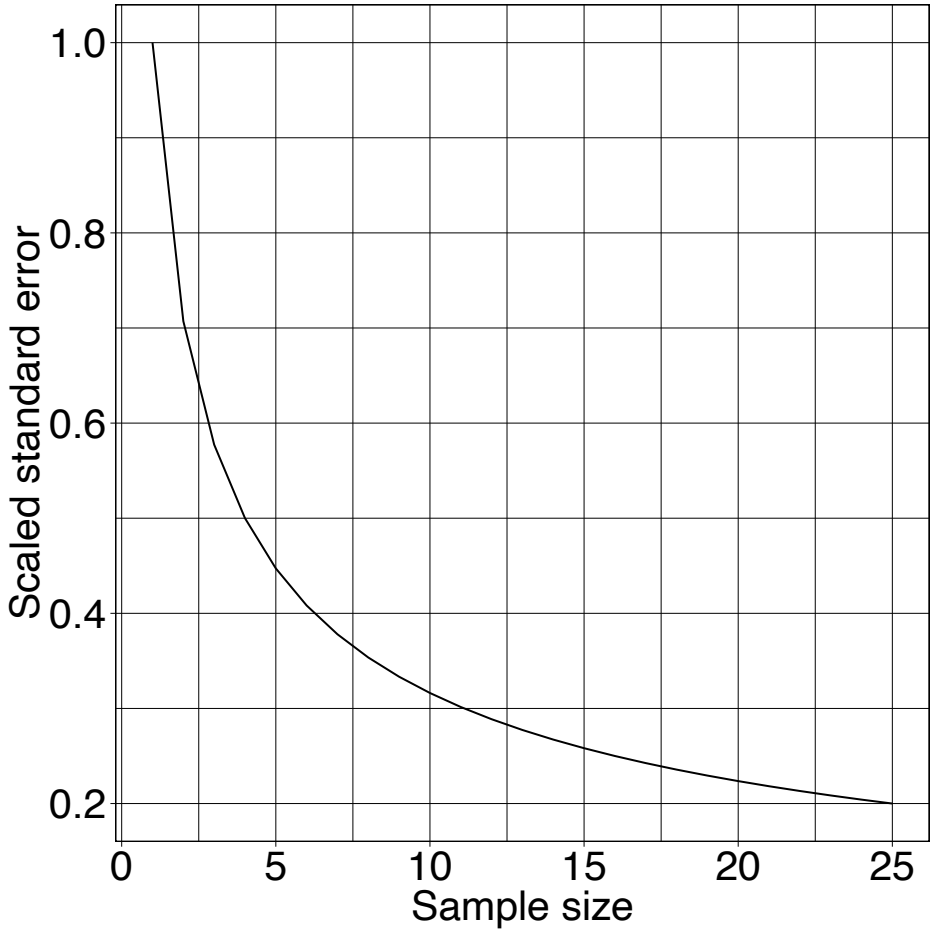


E10.5 - 12.5

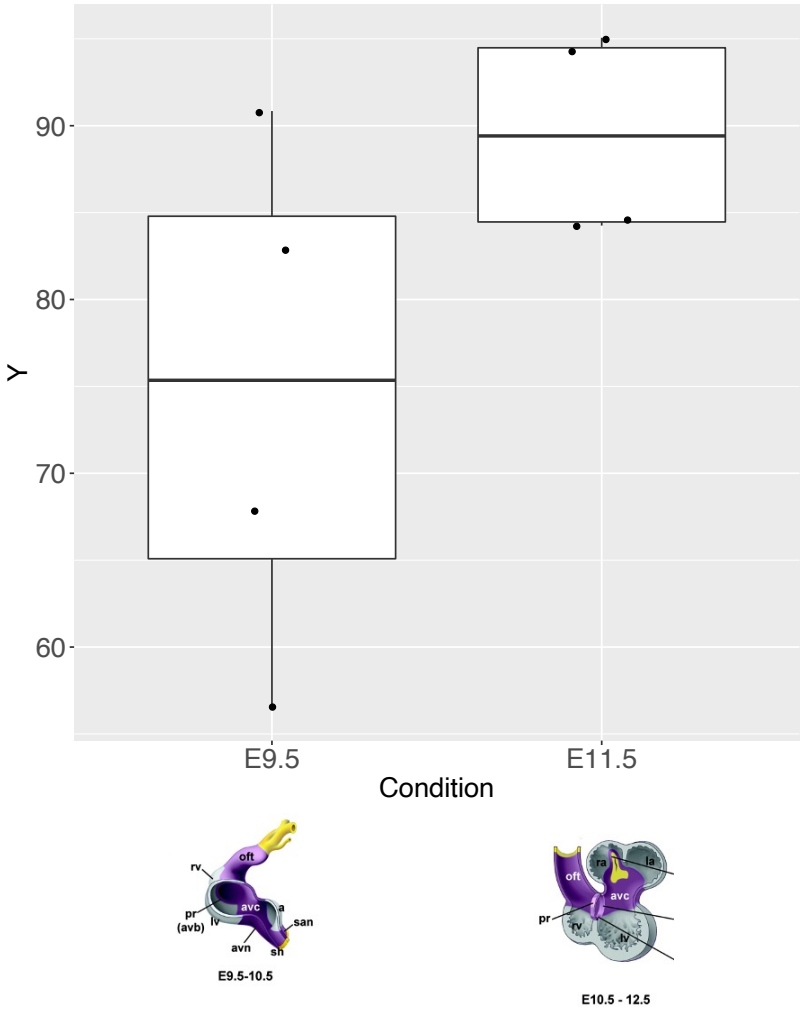
Aggregation: one number to capture an entire distribution



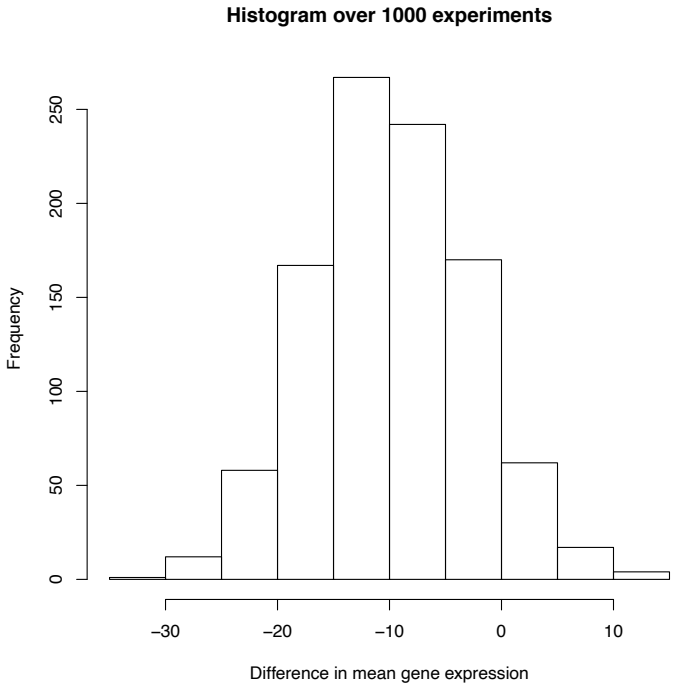
Information on aggregate measure: rate of gain decreases with increasing sample size



Inter-comparison: with limited data can make conclusions applicable to larger target population



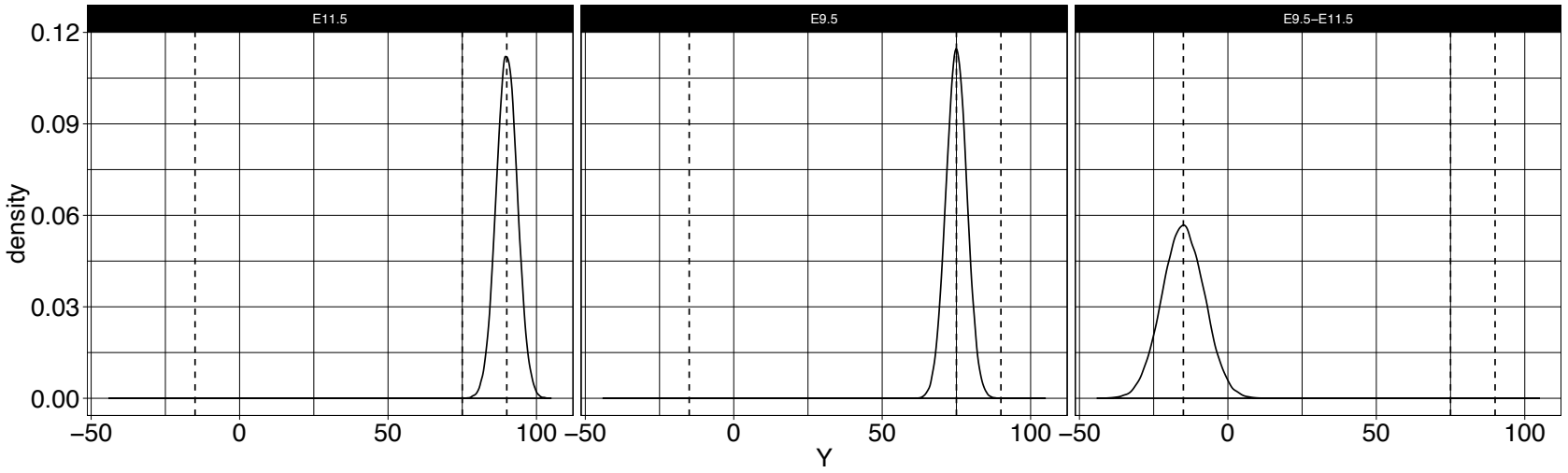
Convince a skeptic: Repeat this experiment 1000 times

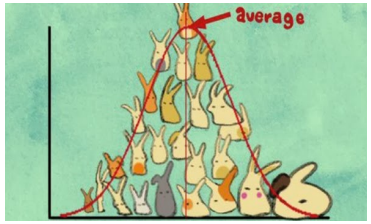


In Science, we can not afford to have millions of samples! The frame of reference for statistical decision making is provided by the sampling distribution of a statistic

Central limit theorem allows us to estimate the variation of the location of the distribution

$$E11.5 : Normal\left(90, \frac{7}{\sqrt{4}}\right) \qquad E9.5 : Normal\left(75, \frac{7}{\sqrt{4}}\right) \qquad E9.5 - E11.5 : Normal\left(75 - 90, \frac{7+7}{\sqrt{4}}\right)$$





The sampling distribution of the mean approaches a normal distribution as the size of the sample increases, regardless of the shape of the original population distribution

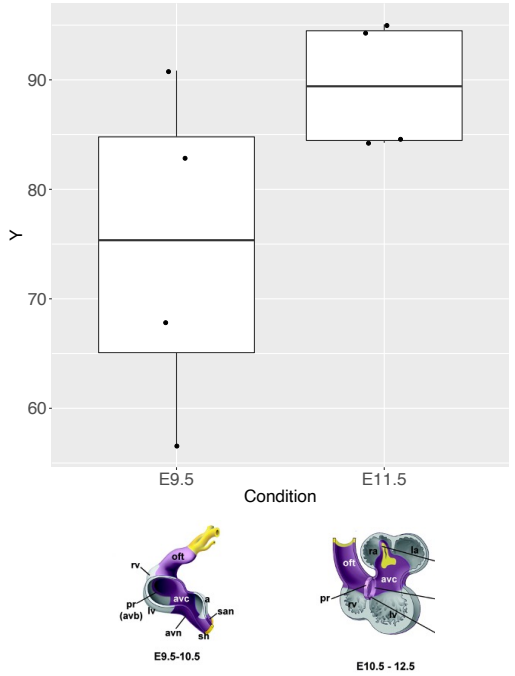
Two conclusions from the data

Conclusion 1: The difference is interesting, biologically meaningful – PI happy, start writing manuscript, plan further experiments.

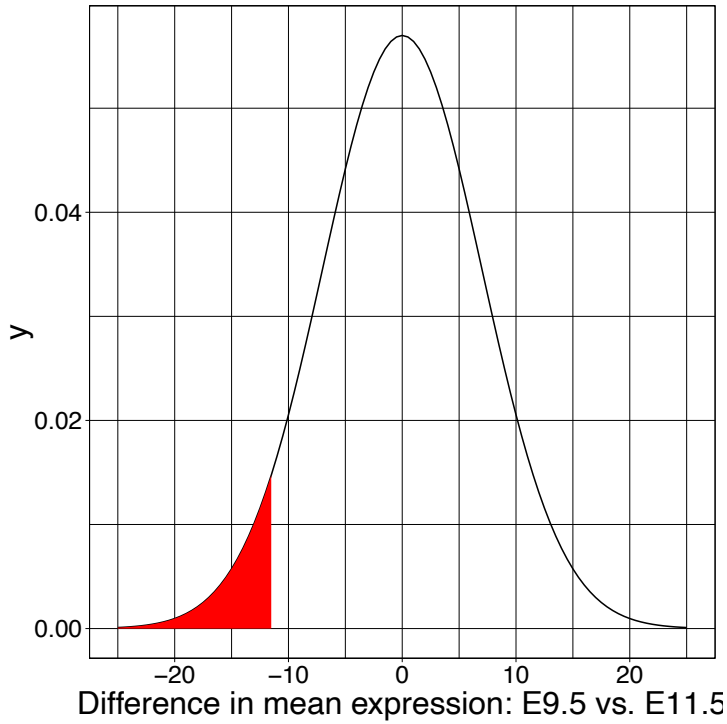
Conclusion 2: Skeptical viewpoint, there is no difference, or unable to conclude that there is one – back to the drawing board.

All statistical hypothesis testing is based on the latter the skeptical viewpoint

How can we establish which of these alternatives has the higher probability of being true?



Theoretical distribution of difference in means under the skeptical view-point/Null hypothesis



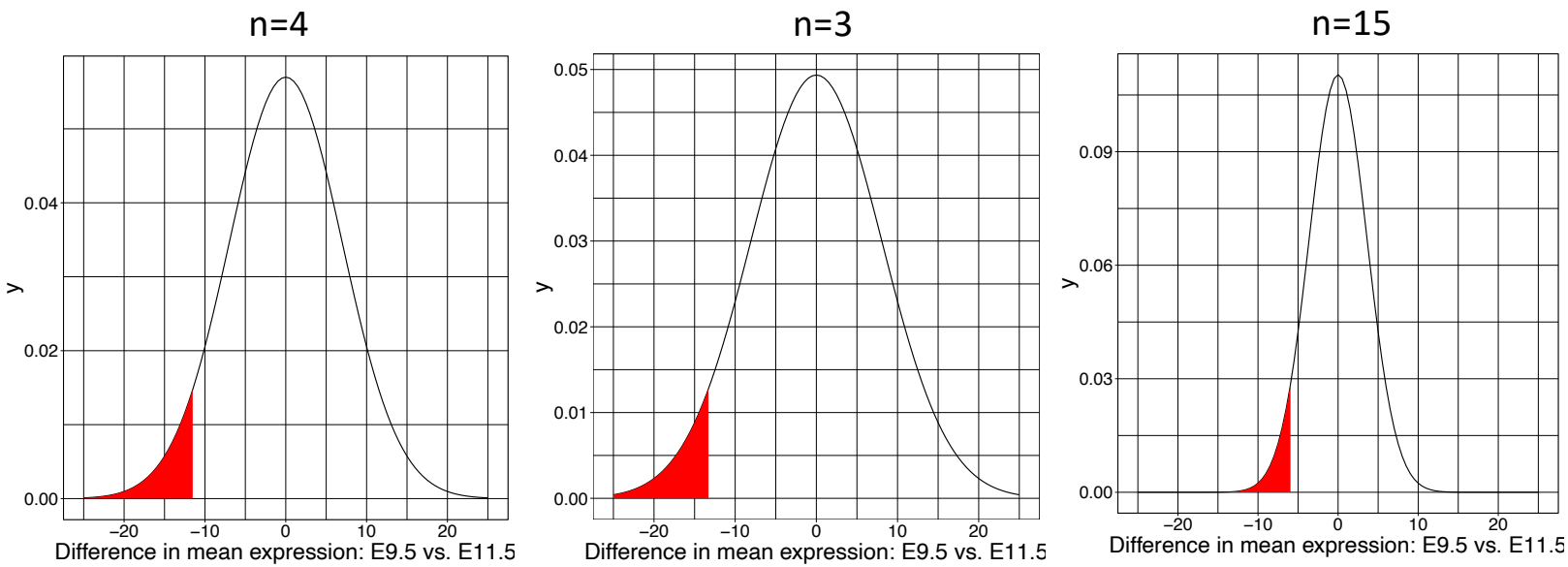
Type I error and p-value

H0 No differences – TRUE

Reject H0 – wrong decision

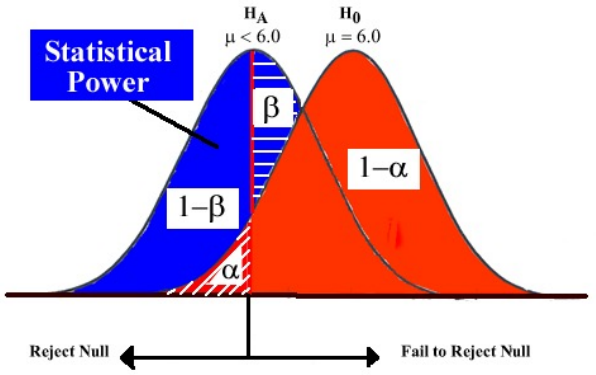
Your findings are significant when in fact they have occurred by chance - FP

Alter the number of replicates



Test errors

		Reality	
		H_0	H_1
Result of the test	Accept H_0	OK	Type 2 error (β)
	Reject H_0	Type 1 error (α)	OK



α = false positive rate

= P(H1|H0)

Most of the time equal to 5%

$$1 - \beta = \text{power of the test}$$

is the likelihood of a significance test detecting an effect when there actually is one

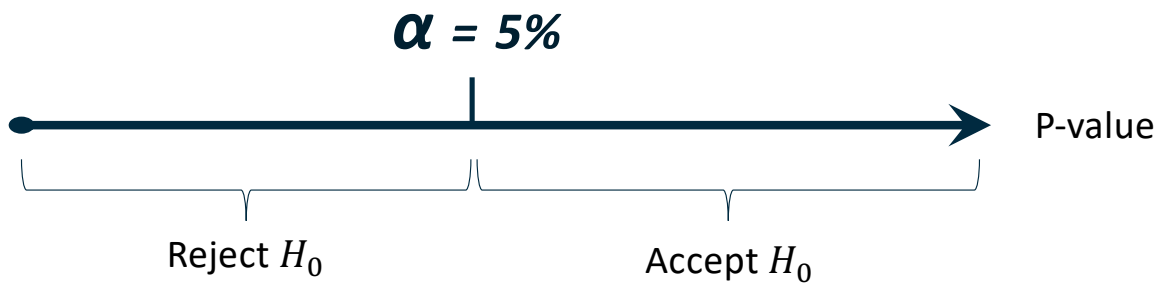
Power is the probability of avoiding a Type II error

P-value

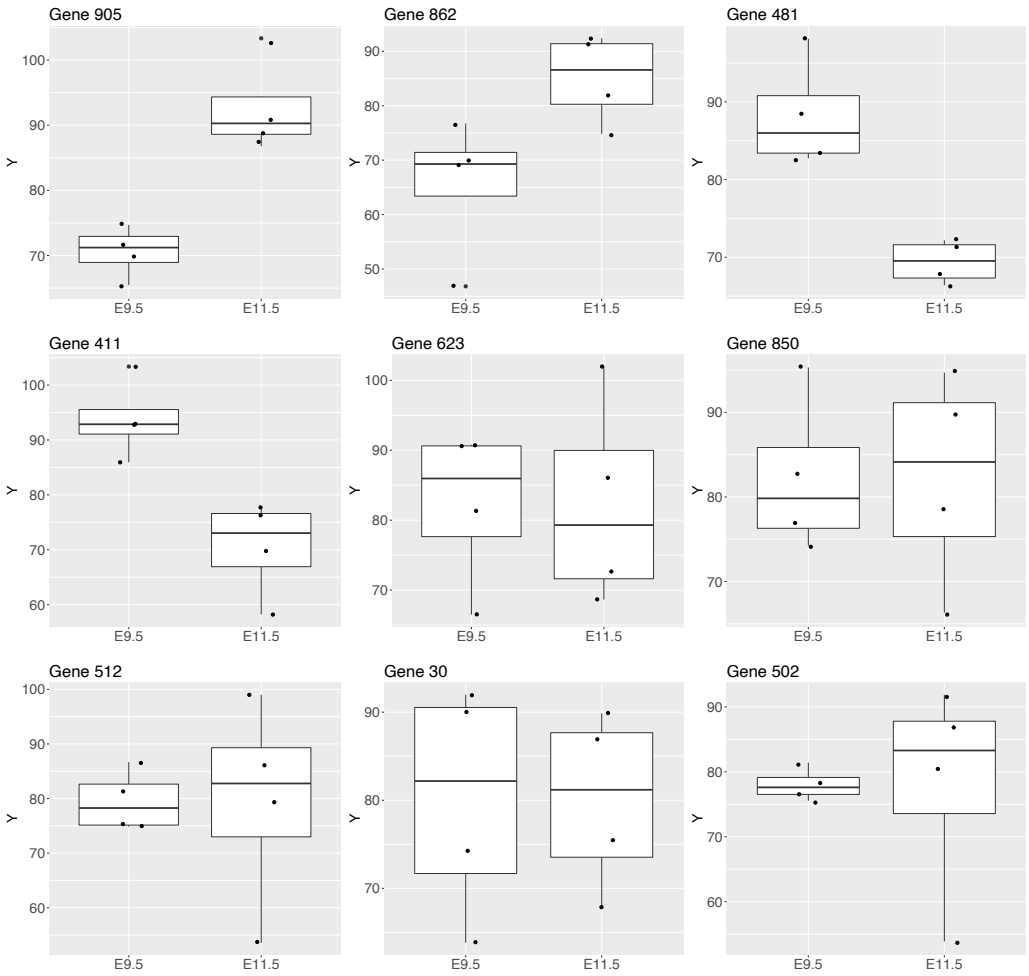
P-value = probability that the result (significance) is only due to chance

P-value = Probability of incorrectly rejecting the null hypothesis, given that it's true (false positive).

Compared to α rate chosen

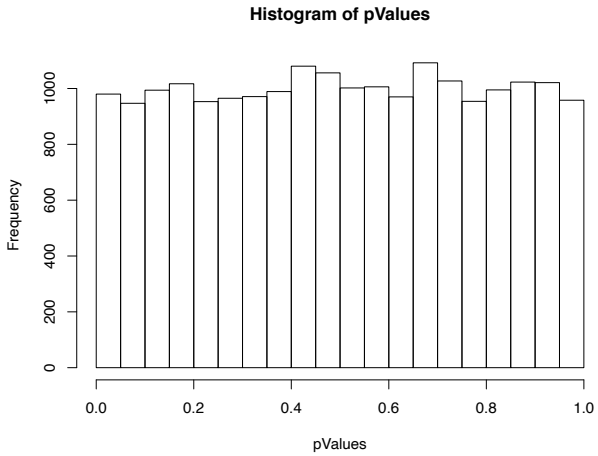


Testing for differences in expression of multiple genes

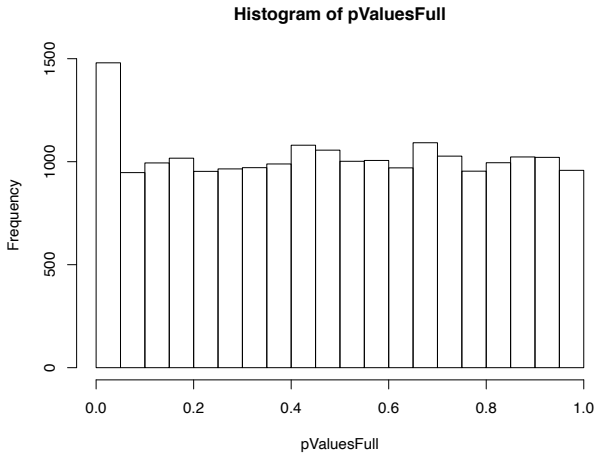


Look at distribution of p-values

No real differences



Possible differences



Multiple testing procedures

If running a lot of test at 5% → Many false positives

Choosing a lower value

Bonferroni method : $5\% / \mathbf{number\ of\ tests}$

Tukey : $5\% / \sqrt{\mathbf{number\ of\ tests}}$

Break ~ 5 minutes

For more details and questions:
michela.traglia@gladstone.ucsf.edu
reuben.thomas@Gladstone.ucsf.edu

Please take the survey:
https://www.surveymonkey.com/r/F75J6VZ

Outline

Experimental design principles

- Types of analytical studies
- Independent variable / response variable
- Target population and generalization

Statistical principles

Experiment: Assess differences between gene expression at two developmental stages

- Aggregation and inter-comparison
- P-value and multiple test correction
- Biological and technical replicates
- Power and sample size
- Overview statistical tests based on response and predictor

Confounding factors

Good vs bad experimental design

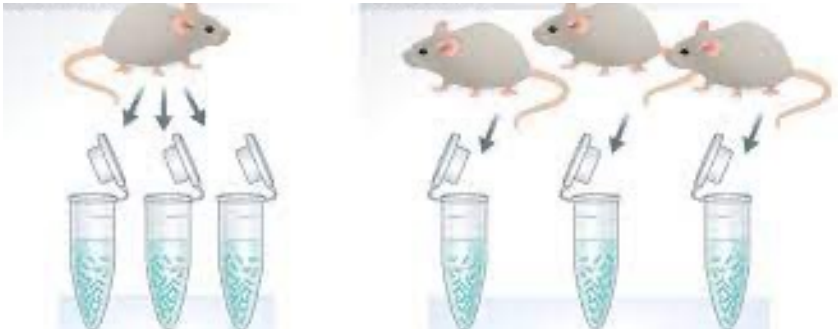
Replicates

- How large differences are you looking for?
- What is the expected expression difference of targeted biology in these samples?
- Will ”no change” be a desired significant result?

Biological vs technical replicates

- Number of replicate runs that will give a high probability of detecting an effect of practical importance
- Use biological replicates to answer biological questions, and technical replicates to answer technical questions

Poll

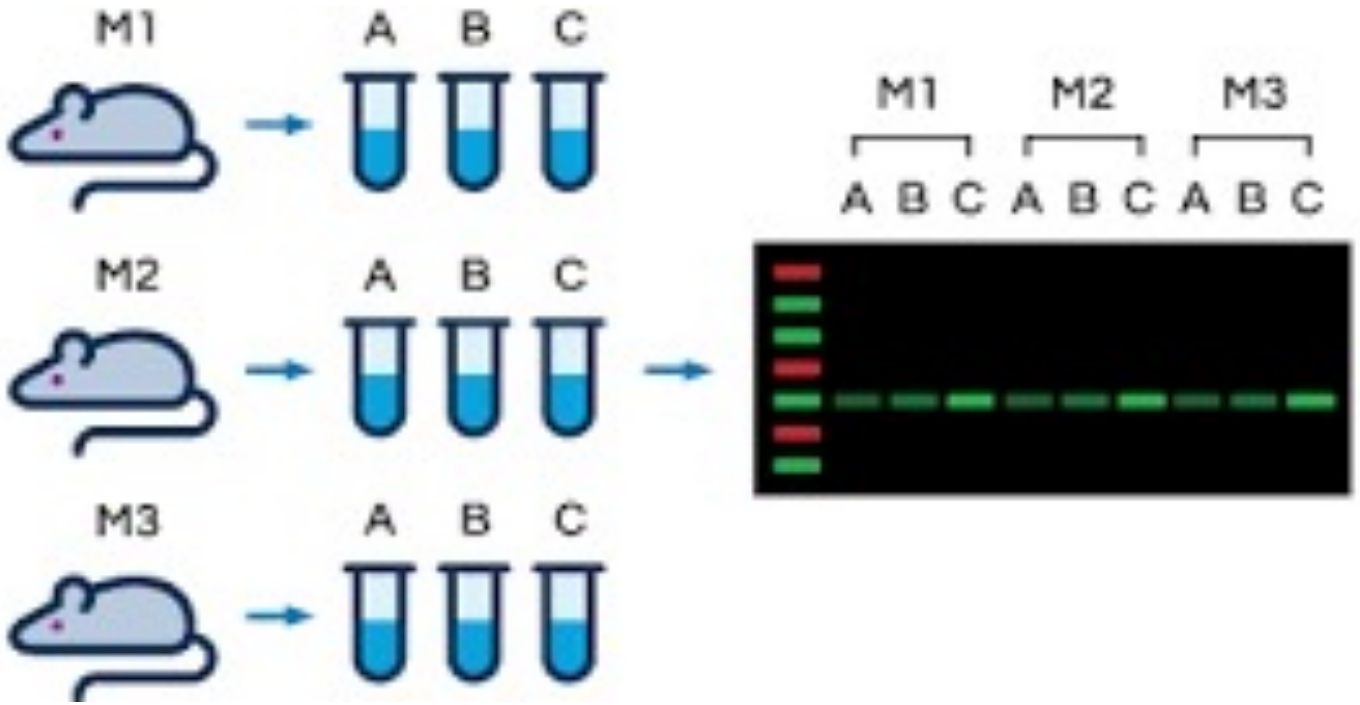


A B

Cell cultures have originated from one mouse (Exp A) and from 3 mice (Exp B). What do you have in experiment B?

1) 3 technical replicates
2) 3 biological replicates

Biological vs technical replicates



Effect size and power

Sample size -> amount of information -> precision (margin of error) / level of confidence in our sample estimates.

- High variability -> Greater uncertainty
- Larger sample size -> more information -> less uncertainty

Power – probability that we find statistically significant evidence of a difference between the groups, given that there is a difference in the population.

Effect size is the estimated difference between the groups that we observe in our sample

Small effect size -> large sample size to detect the difference OR effect masked by the randomness

Is a larger sample sizes always better?

+ Greater precision and power
- Cost more time and money

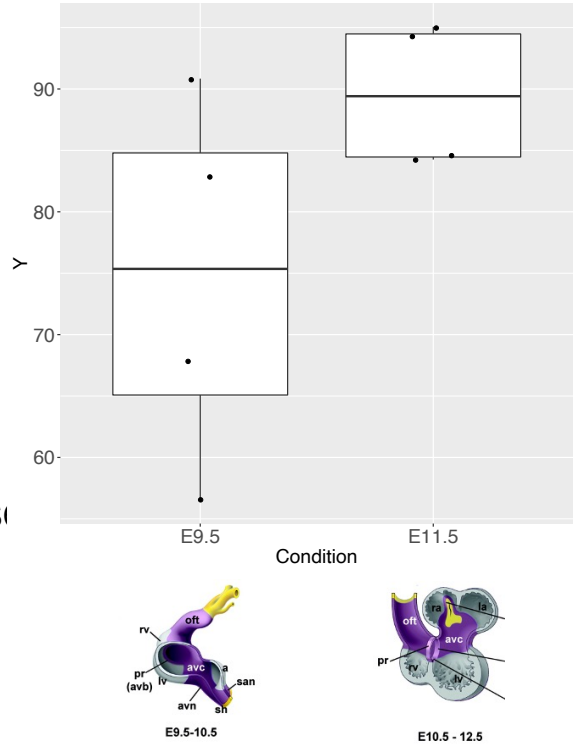
The goal is to collect enough data from a sample to statistically test whether you can reasonably reject the null hypothesis in favor of the alternative hypothesis

Perform a sample size calculation before – acceptable power >0.8

If there are true effects to be found in 100 different studies with 80% power, only 80 out of 100 statistical tests will actually detect them

How many n?

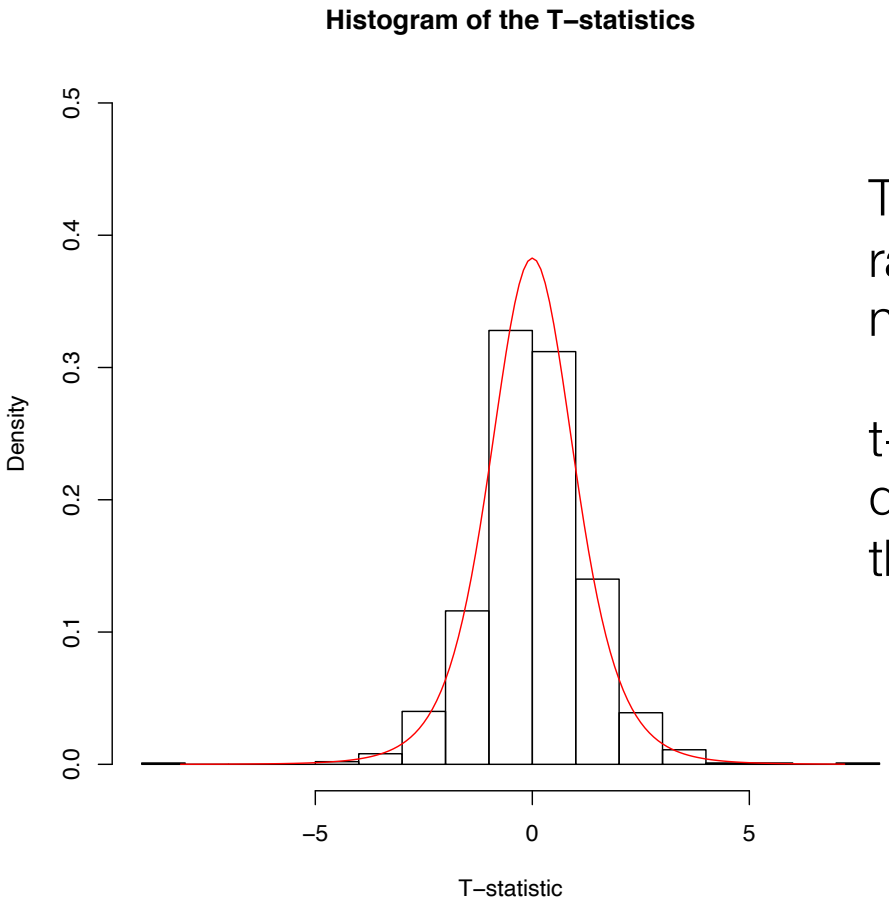
- Identify parameters of interest given experimental design – two variables models to more complex multivariate designs
- Test statistic for the parameters of interest
- Decide on the number of samples (to be drawn from the population of samples) under each setting
- Estimates n of variation and correlation between variables of interest – us pilot data or publicly available data
- Sampling distribution of this test statistic
- Check assumptions for the validity of the sampling distributions



Z/T-statistic

$$Z = \frac{mean\left(Y_{E9.5}\right) - mean\left(Y_{E11.5}\right)}{sd(Y)\sqrt{\frac{1}{n} + \frac{1}{n}}}$$

T-statistic and sampling distribution



T-distributions assume that you draw repeated random samples from a population where the null hypothesis is true.

t-value from your study in the t-distribution to determine how consistent your results are with the null hypothesis.

Outline

Experimental design principles

- Types of analytical studies
- Independent variable / response variable
- Target population and generalization

Statistical principles

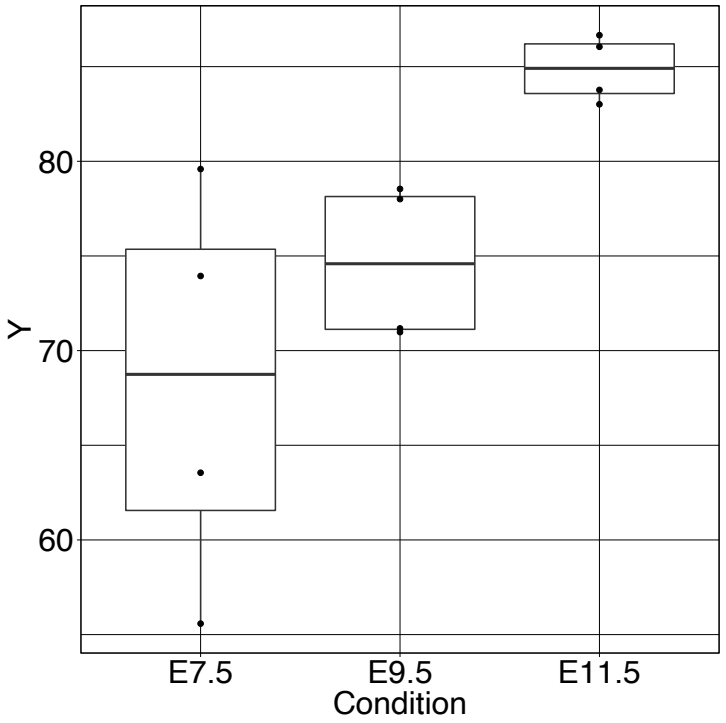
Experiment: Assess differences between gene expression at two developmental stages

- Aggregation and inter-comparison
- P-value and multiple test correction
- Biological and technical replicates
- Power and sample size
- Overview statistical tests based on response and predictor

Confounding factors

Good vs bad experimental design

Continuous response and categorical predictor



Three unrelated groups

Differences among means

Y: gene expression
X: development time

One-way ANOVA – F-statistics

Two categorical variables

	In TGF-b signaling pathway	Not in TGF-b signaling pathway
Differentially expressed	20	980
Not differential expressed	80	18920

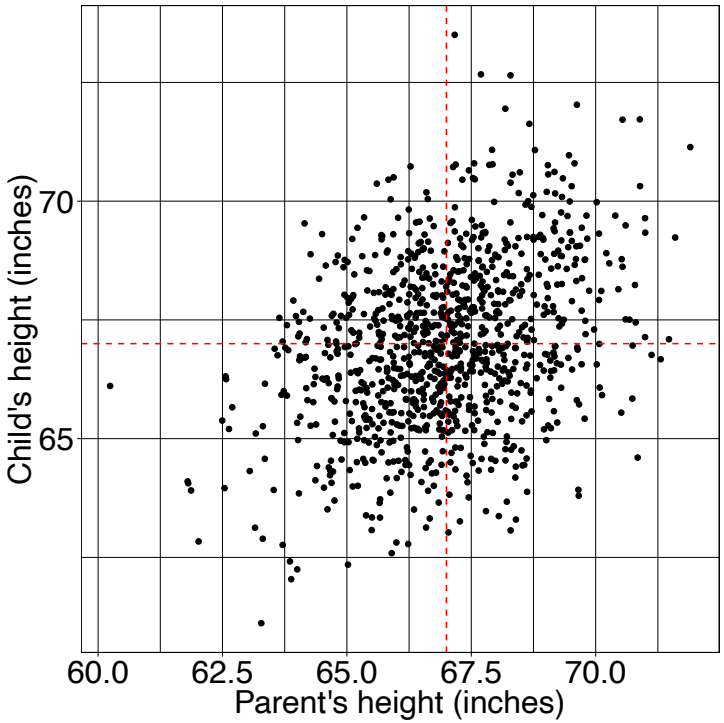
Y1: gene differentially expressed or not
Y2: gene in TGF-b signaling pathway or not

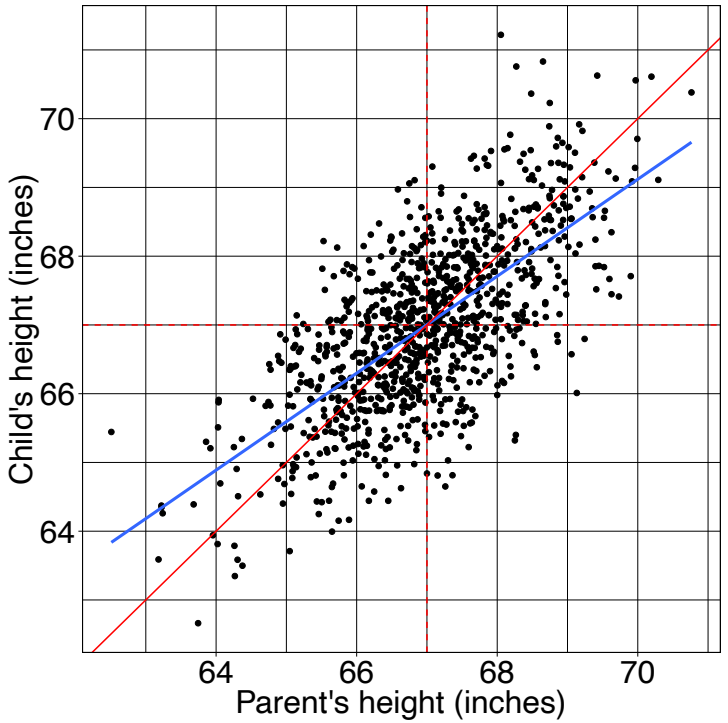
Odds ratio, Chi-square statistics

The chi-squared test works by calculating the frequencies we would expect to see in the cells if there were absolutely no association

Continuous response against a continuous variable

Associate multiple (noisy) factors with each other



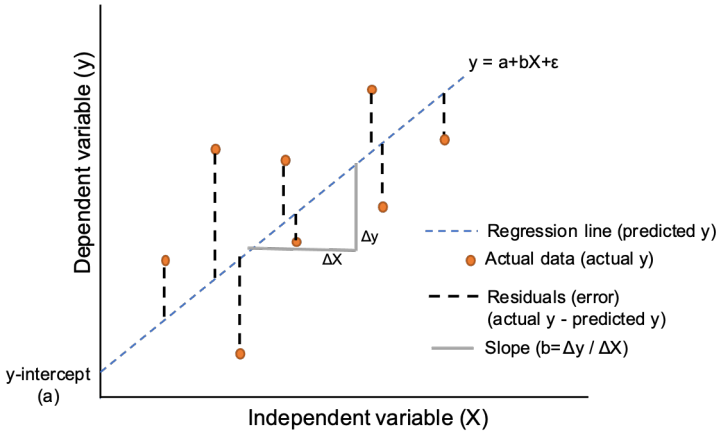


Y: Child's height

X: Parent's height

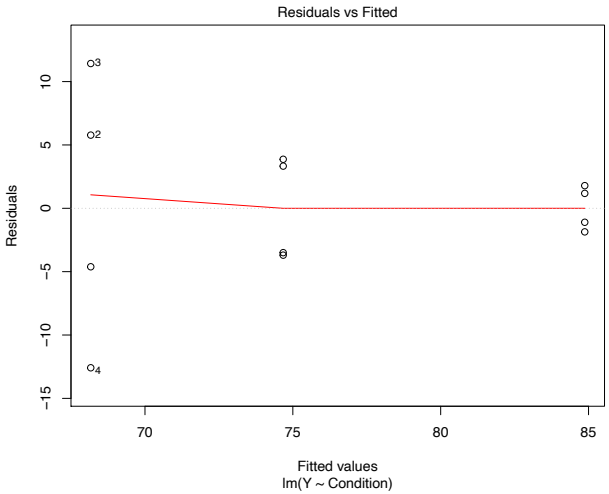
Slope, linear regression

Residual: Variation left over after we have captured the known effects



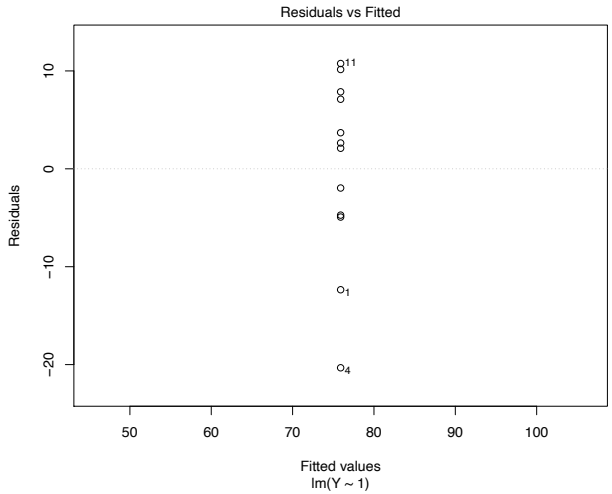
Full model

lm (Y ~ condition)



Mean model

lm (Y ~ 1)



Residual: Predicted - Observed

Outline

Experimental design principles

- Types of analytical studies
- Independent variable / response variable
- Target population and generalization

Statistical principles

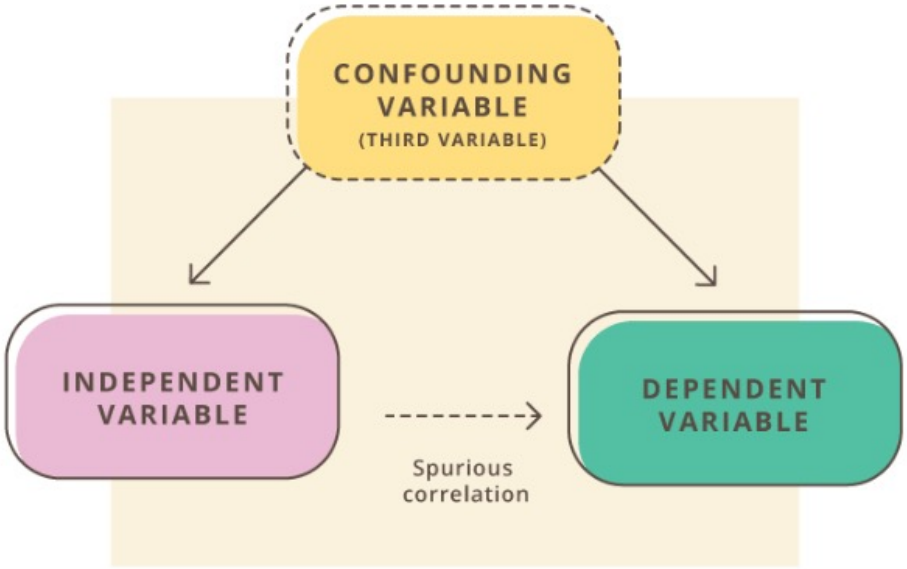
Experiment: Assess differences between gene expression at two developmental stages

- Aggregation and inter-comparison
- P-value and multiple test correction
- Biological and technical replicates
- Power and sample size
- Overview statistical tests based on response and predictor

Confounding factors

Good vs bad experimental design

A third variable



What at first looks like a causal relationship between IV and DV is ultimately spurious. The confounding variable is the hidden explanation.

Be aware of confounding factors

In observational studies (case-control, cohort studies):

- lurking variables could cause unusual interpretations of data and the relationships between variables

In experimental studies:

- design the experiment to eliminate (as much as possible) the risk of lurking variables

Which potential variables could be affecting the relationship between the variables in your study?

Chemical in pregnant maternal blood correlating with healthy children

Mother with high levels of chemicals blood

->

Correlation

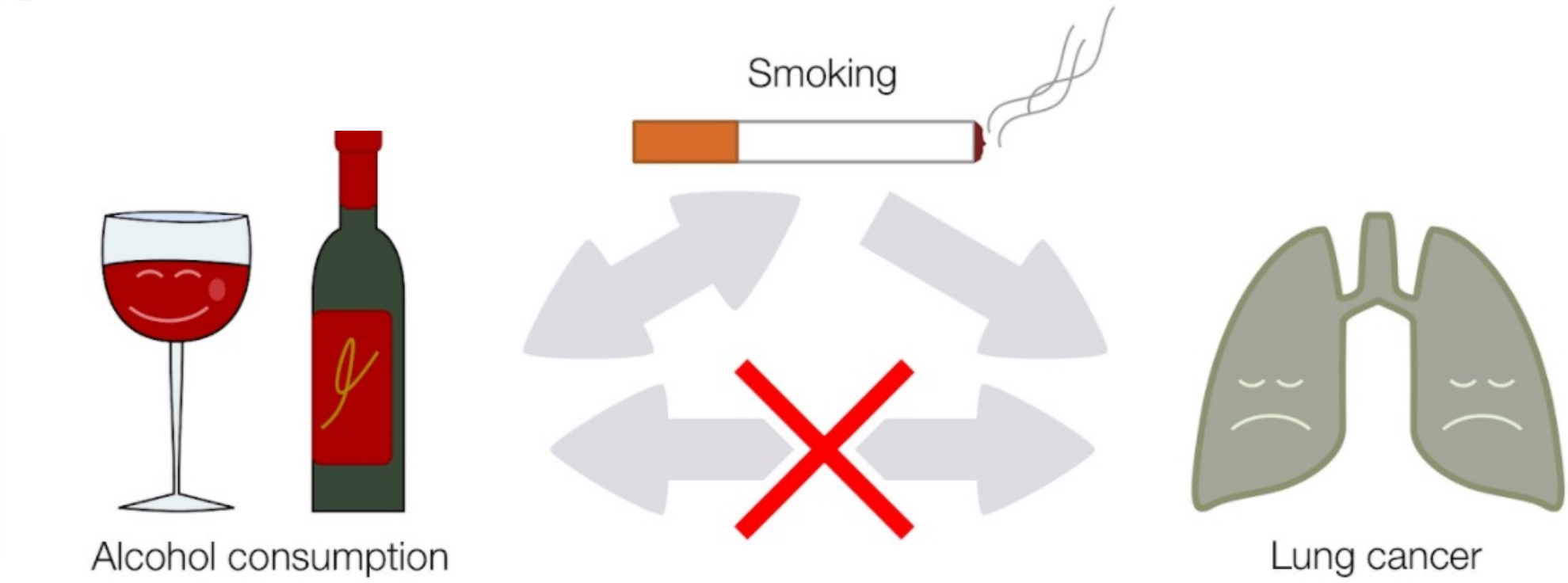
Healthy kids

High chemicals levels – not developing autism
Low chemicals levels – children developing autism

?

Genetics make-up controlling metabolism of chemical

Correlation vs Causation



Poll

Research question: *Does light exposure improve learning ability in mice?*

What can be the source of variability, confounding the outcome?
Select all that apply

1. Mouse inbred strains
2. Genetic background
3. Learning environment
4. All are ‘independent variables’

Factors potentially affecting the response

Biological factors that could affect the response

- -BMI
- -gender

Non-biological or technical factors that could affect the response.

- -time/day/month of experiment/batch
- -reagents used, reagents batch used
- -technician

Capture effects of interest and avoid unwanted variation in experiment

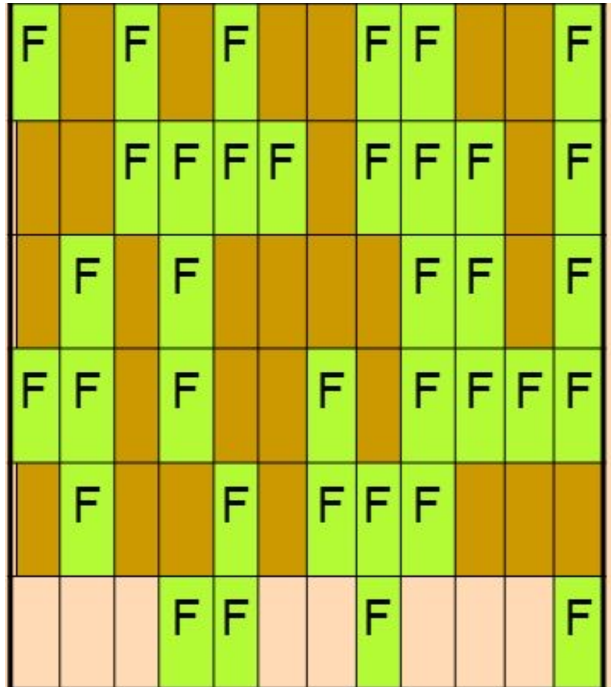
- ✓ Identify the response and variables of interest
- ✓ Identify target population that you want to base your claims on
- ✓ Identify factors that affect the response of interest
- ✓ Choose samples from target population

Randomly assign samples across different levels of factors affecting response

Block out variation that is not of interest by randomly assigning to levels of factors within a block

Randomization

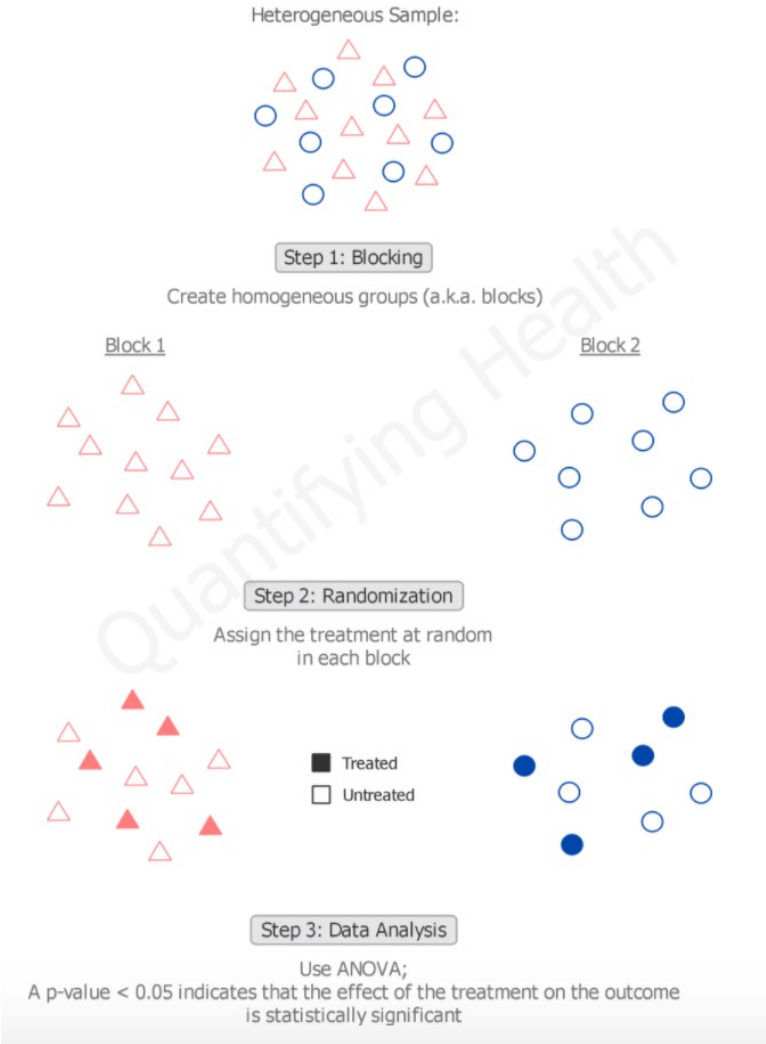
Fisher -> helps to avoid confusion or biases due to changes in background or lurking variables.



One of the main purposes for experimental designs is to minimize the effect of experimental error.

Randomization, replication, and blocking, are methods of error control.

Randomized block design

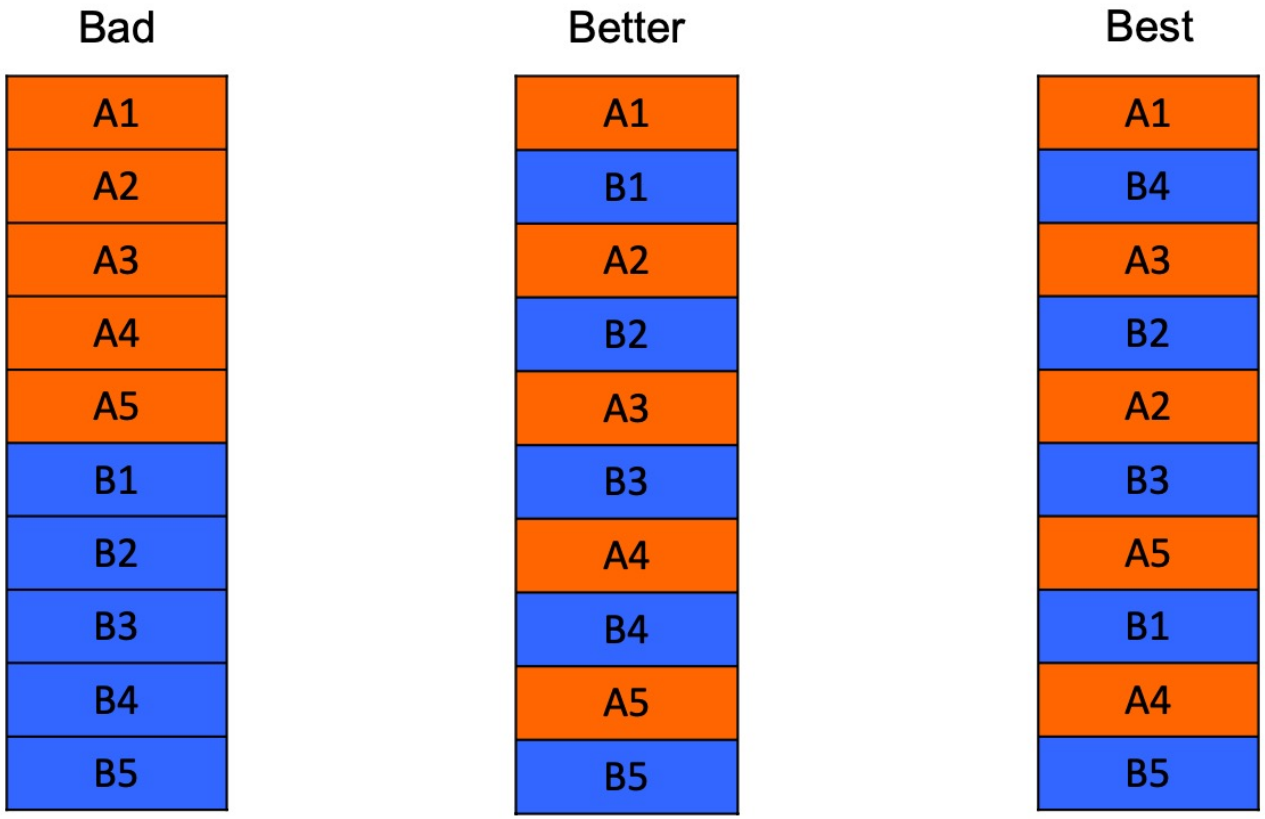


Treatment	
Placebo	Vaccine
500	500

Gender	Treatment	
	Placebo	Vaccine
Male	250	250
Female	250	250

https://quantifyinghealth.com/randomized-block-design/
https://stattrek.com/experiments/experimental-design.aspx

Good vs Bad experimental design



- The default analysis assumes the data have come from a completely randomized design.
- In practice, this is often a false assumption.

Poll

	Design 1 – Sample prep date	Design 2 – Sample prep date
Sample_1_E9.5	Jan 9th, 2019	Jan 11th, 2019
Sample_2_E9.5	Jan 9th, 2019	Jan 9th, 2019
Sample_3_E9.5	Jan 9th, 2019	Jan 11th, 2019
Sample_4_E9.5	Jan 9th, 2019	Jan 9th, 2019
Sample_1_E11.5	Jan 11th, 2019	Jan 11th, 2019
Sample_2_E11.5	Jan 11th, 2019	Jan 9th, 2019
Sample_3_E11.5	Jan 11th, 2019	Jan 11th, 2019
Sample_4_E11.5	Jan 11th, 2019	Jan 9th, 2019

Which is better design?

1) Design 1
2) Design 2

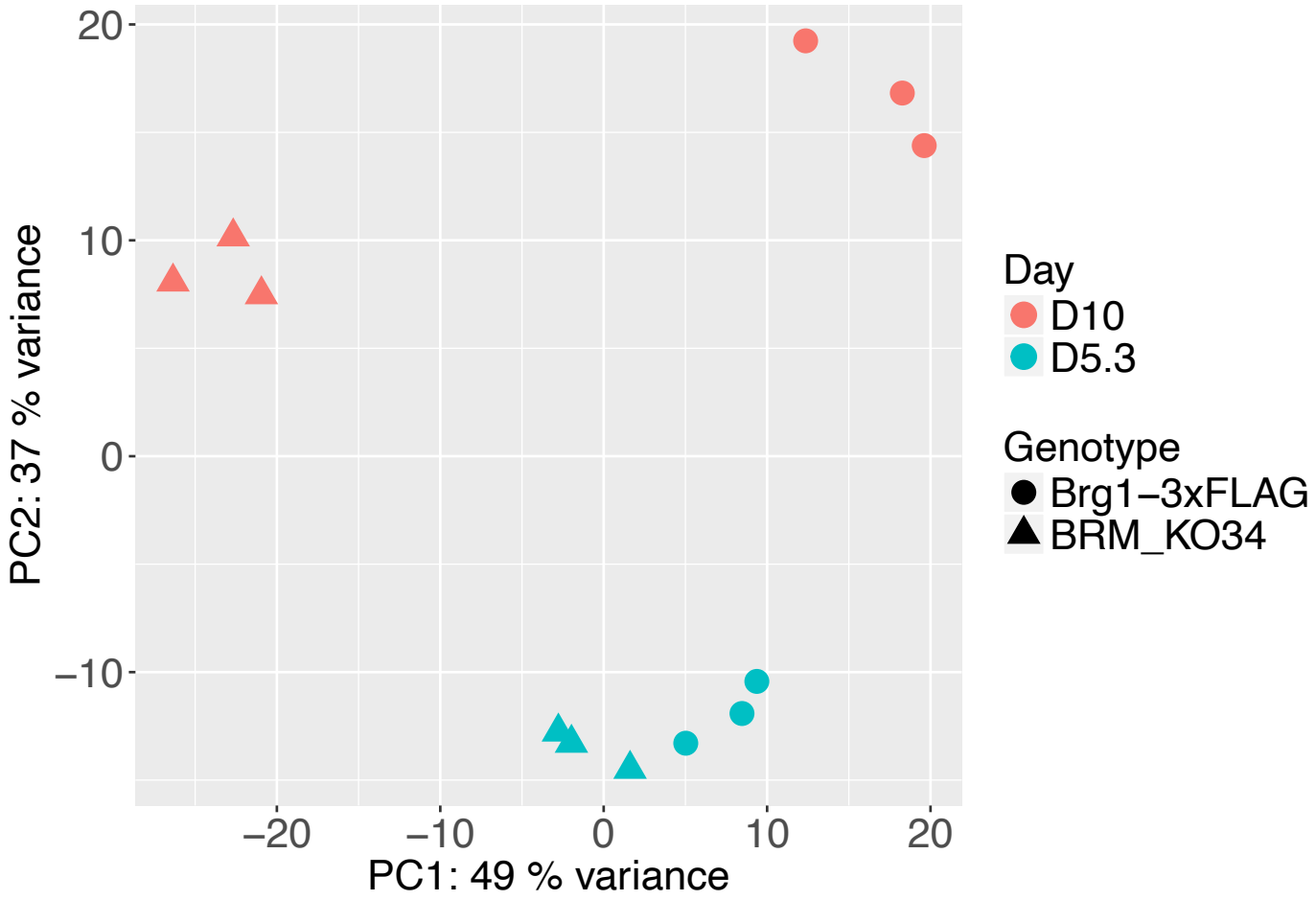
Poll

	Design 1 – Gender	Design 2 - Gender
Sample_1_E9.5	Male	Male
Sample_2_E9.5	Male	Female
Sample_3_E9.5	Male	Male
Sample_4_E9.5	Male	Female
Sample_1_E11.5	Female	Male
Sample_2_E11.5	Female	Female
Sample_3_E11.5	Female	Male
Sample_4_E11.5	Female	Female

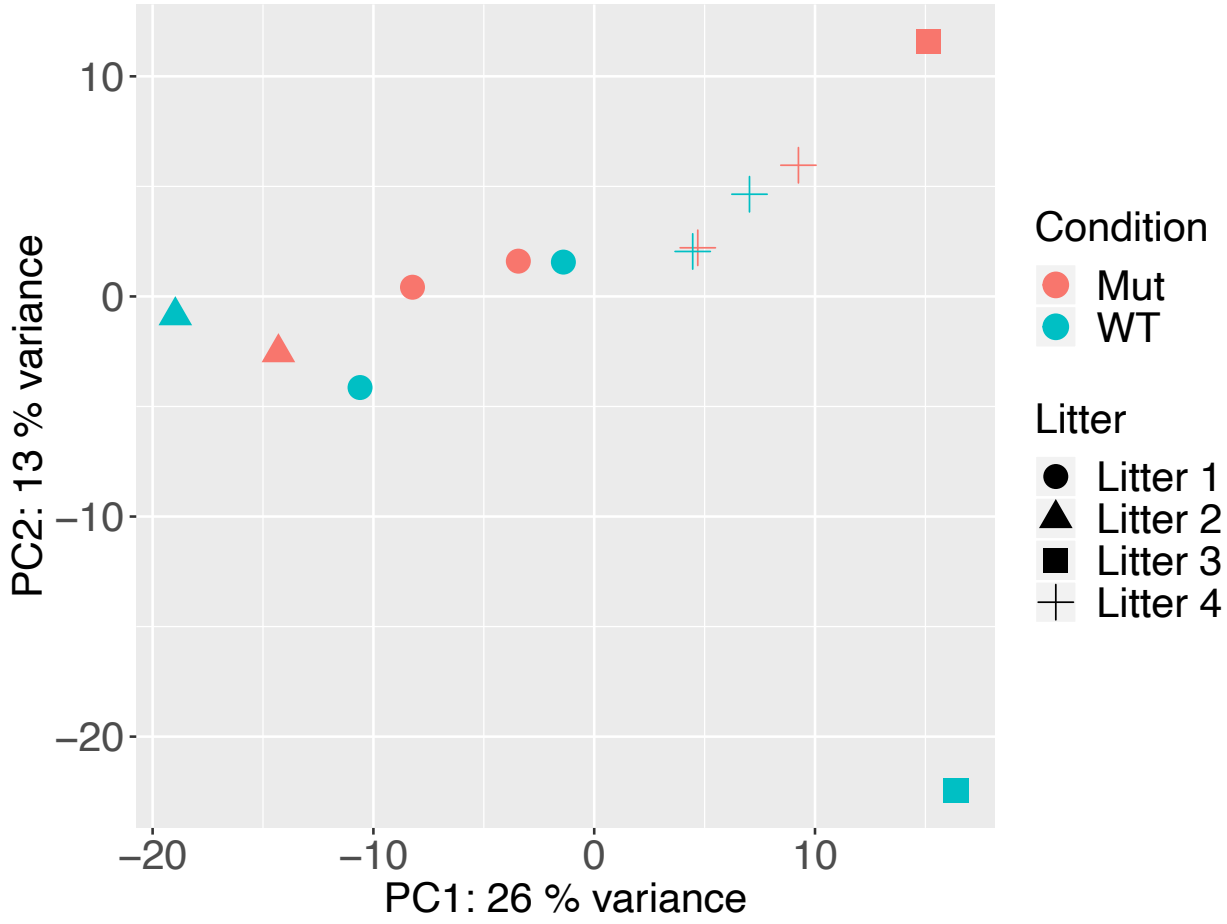
Which is better design?

Design 1
Design 2

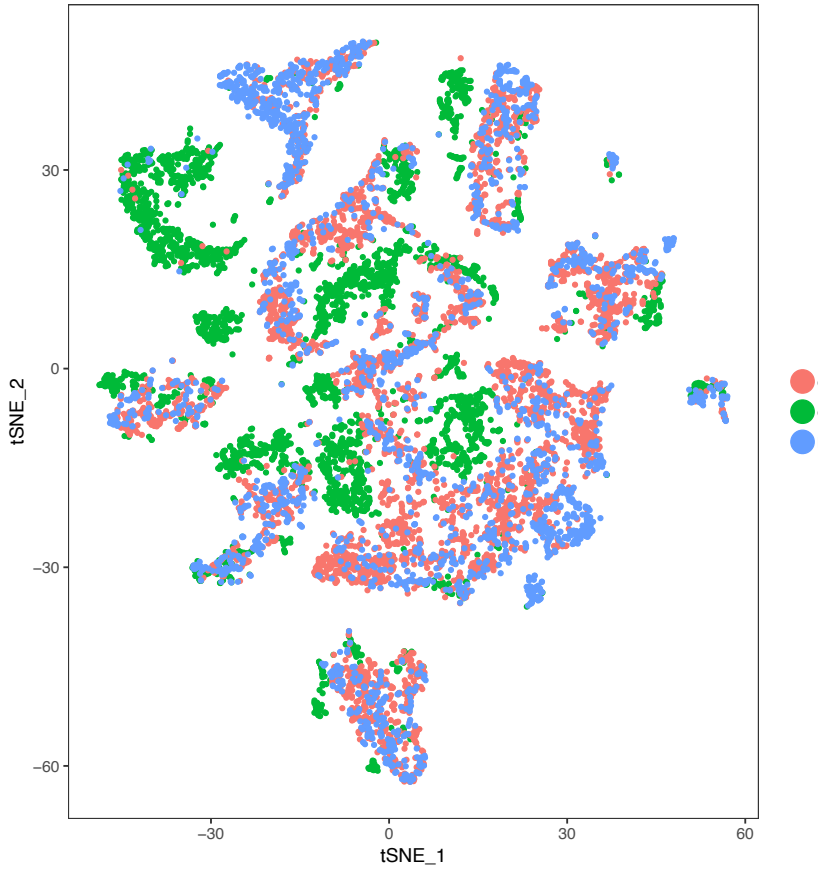
Genotype and development time effect on gene expression



Litter effect dominates the variation

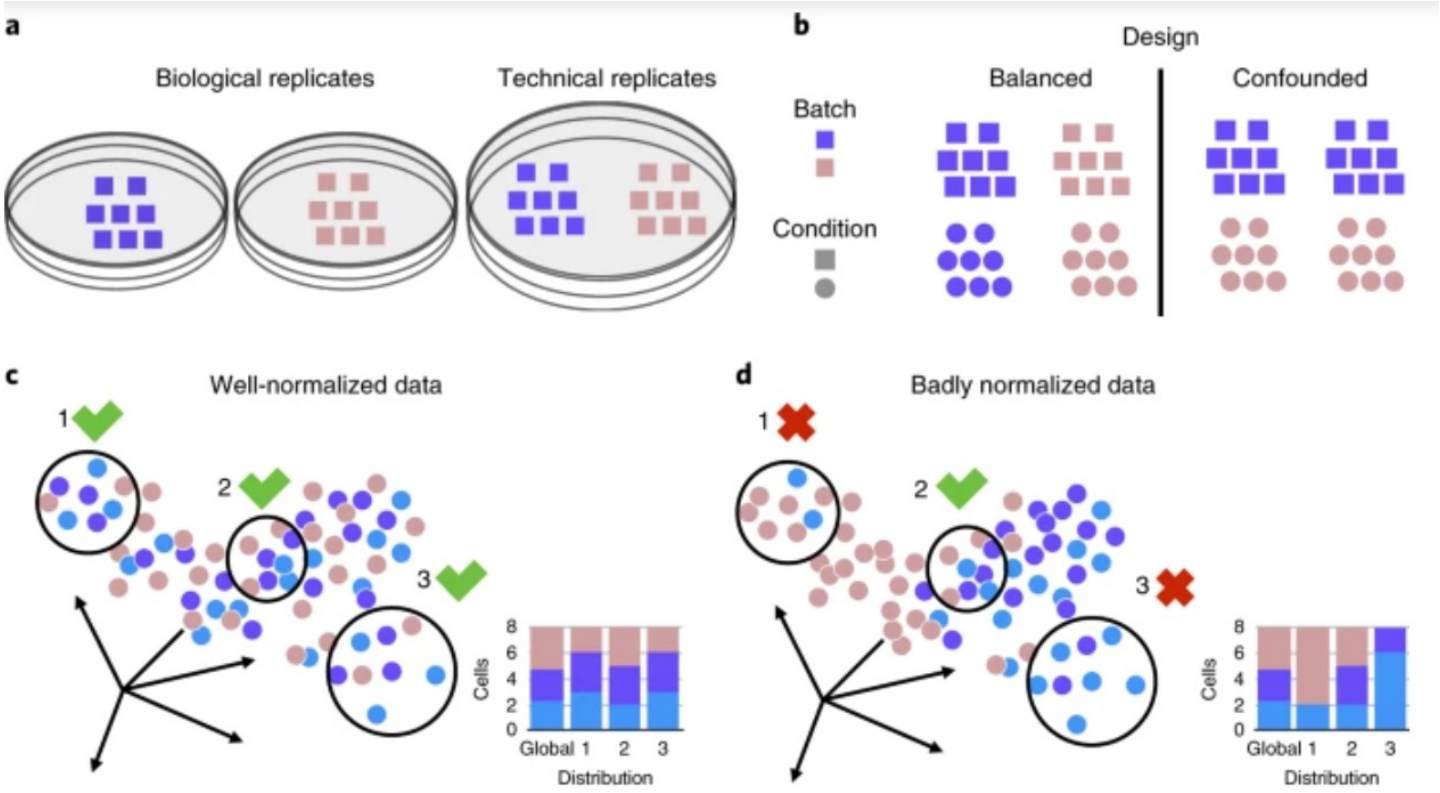


Confounding in scRNA-seq data is a big problem



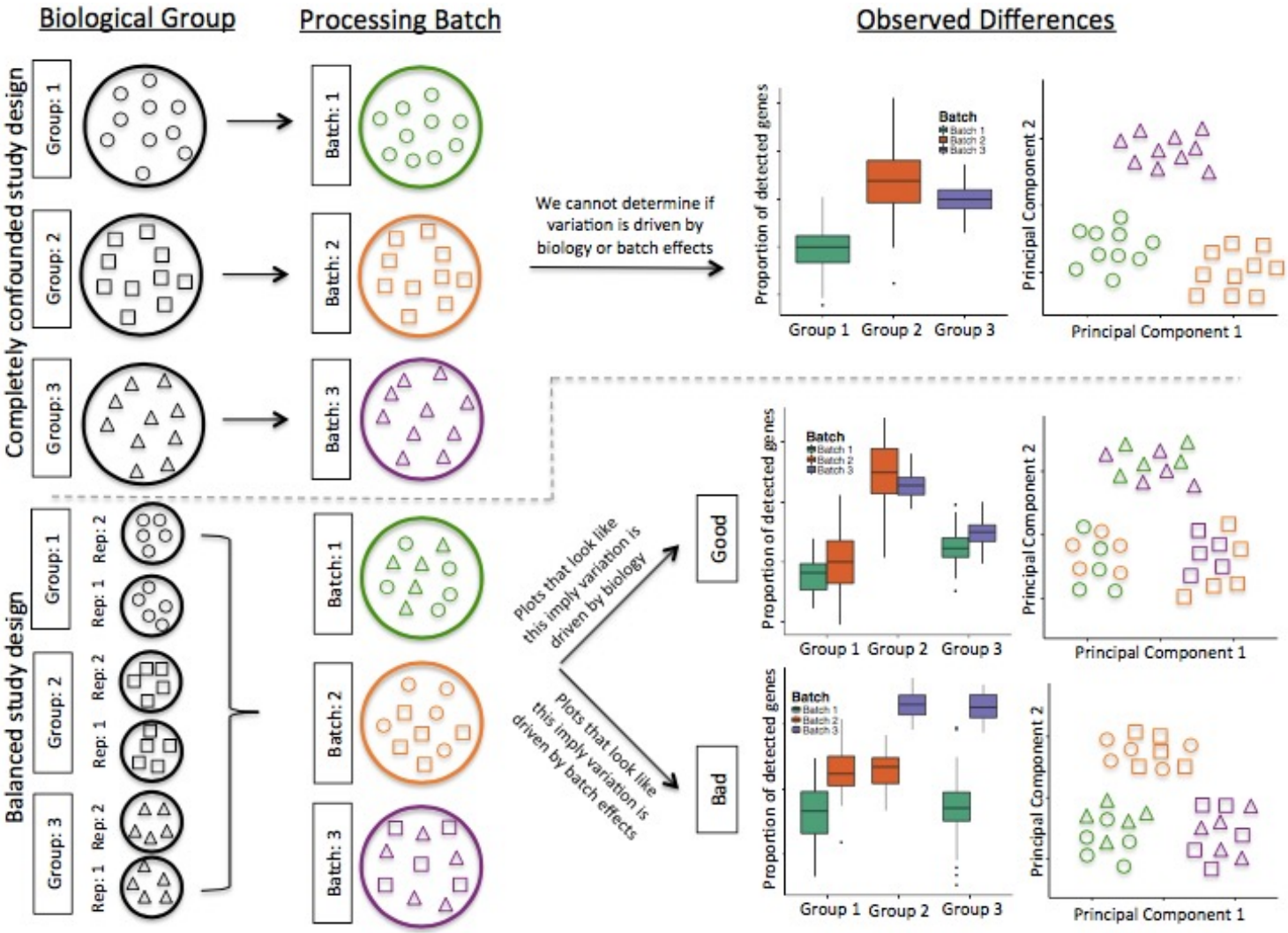
Hicks, S. C., Townes, F. W., Teng, M. & Irizarry, R. A. Missing data and technical variability in single-cell RNA-sequencing experiments. Preprint available from:https://doi.org/10.1093/biostatistics/kxx053 (2017).

Well and badly analyzed scRNA-seq data



https://www.nature.com/articles/s41592-018-0254-1

Confounding biological variation and batch effects



Experimental design allow you to:

(1) choose an experimental design that is appropriate for the research problem at hand

(2) construct the design (performing randomization and determining the number of replicates)

(3) execute the plan to collect the data (or advise a colleague on how to do it)

(4) determine the model appropriate for the data

(5) fit the model to the data

(6) interpret the data and present the results in a meaningful way to answer the research question

Take – home messages

- Plan ahead
- Prevent bias from uncontrollable
- Randomization and balancing
- Write it down in an Experimental plan
- Follow the experimental plan
- Be careful with interpretation of results!

Upcoming workshop at Gladstone:

Fall series

December 3 | Intermediate R: Data Visualization
December 6-7 | Intermediate RNA-Seq analysis Using R

Data Science Training Program

Please visit the link before spring 2022 for new workshops

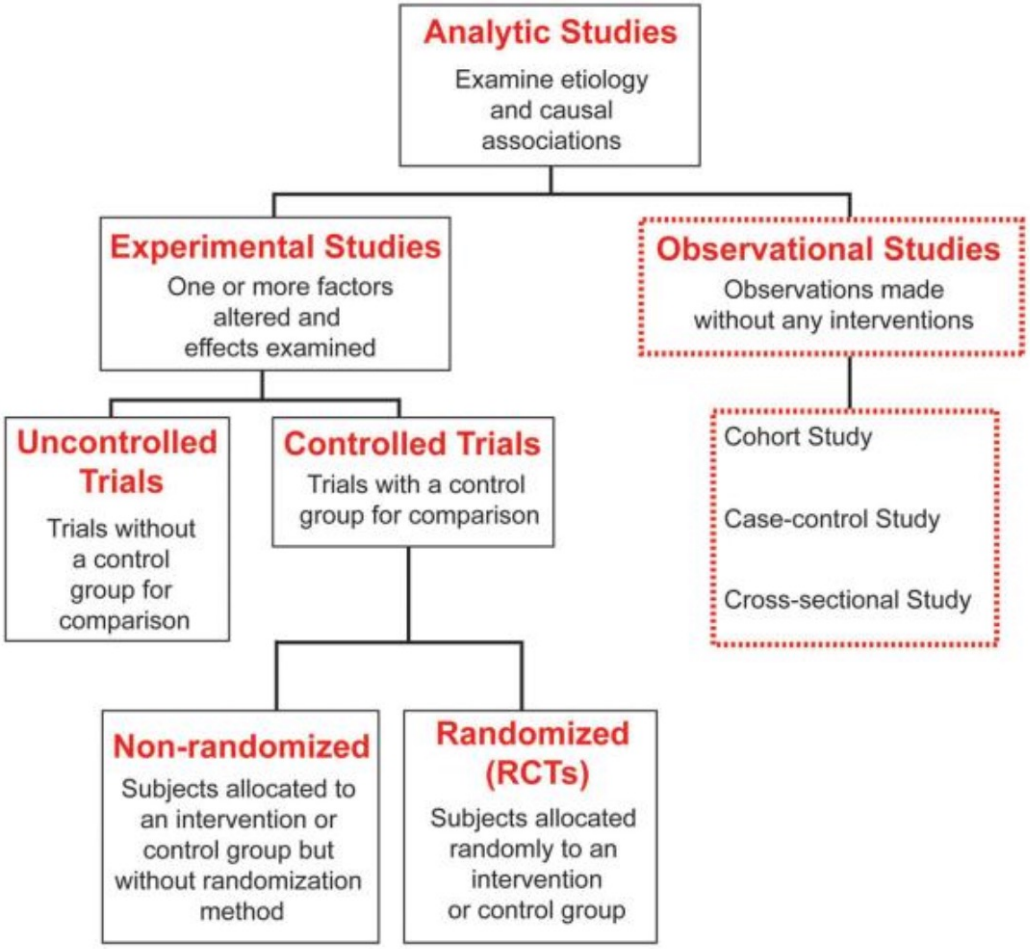
For questions:
michela.traglia@gladstone.ucsf.edu
reuben.thomas@Gladstone.ucsf.edu

Thank you!

Please take the survey:
https://www.surveymonkey.com/r/F75J6VZ

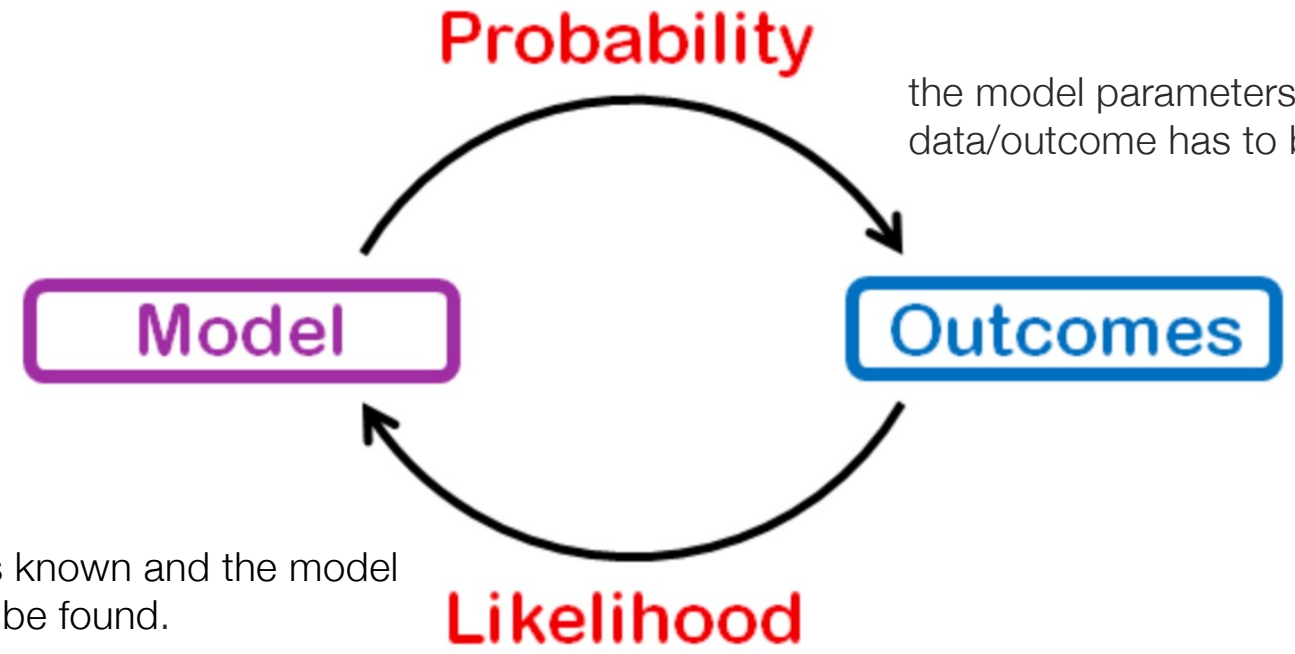
Additional resources

Observational vs Experimental studies



Analytic Study Designs. Adapted with permission from Joseph Eisenberg, Ph.D.

Probability and likelihood



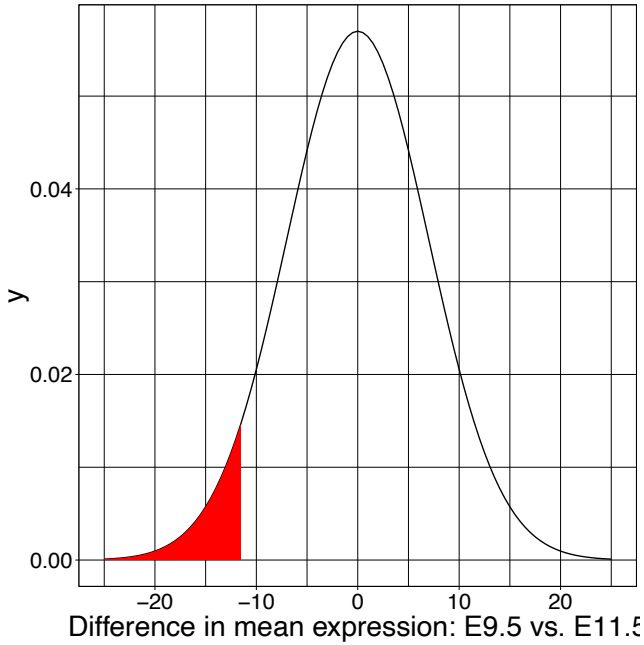
the model parameters are known and the data/outcome has to be found.

the data/outcome is known and the model parameters have to be found.

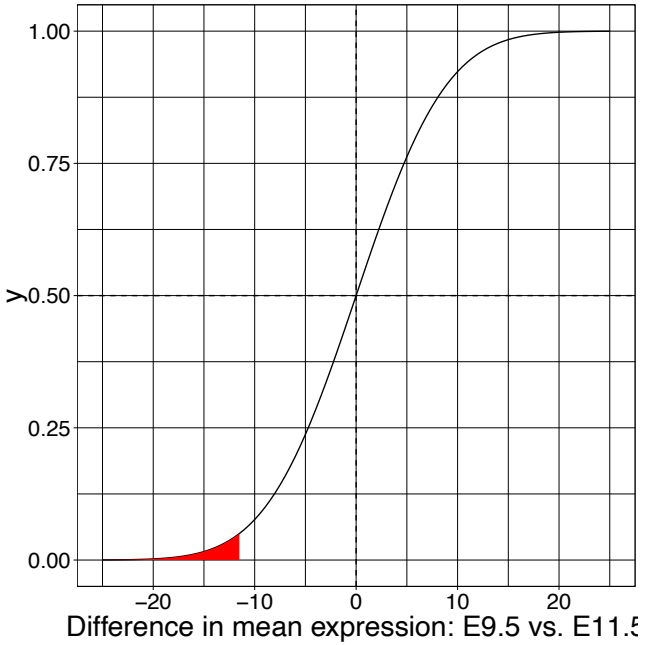
Probability is used to finding the chance of occurrence of a particular situation, whereas Likelihood is used to generally maximizing the chances of a particular situation to occur.

Probability density and cumulative distribution

Density



Distribution



How to choose the right statistical test?